

Nonparametric Bayesian Storyline Detection from Microtexts

Vinodh Krishnan

Georgia Institute of Technology
Atlanta, GA 30308
krishnan.vinodh@gmail.com

Jacob Eisenstein

Georgia Institute of Technology
Atlanta, GA 30308
jacobe@gmail.com

Abstract

News events and social media are composed of evolving storylines, which capture public attention for a limited period of time. Identifying storylines requires integrating temporal and linguistic information, and prior work takes a largely heuristic approach. We present a novel online non-parametric Bayesian framework for storyline detection, using the distance-dependent Chinese Restaurant Process (dd-CRP). To ensure efficient linear-time inference, we employ a fixed-lag Gibbs sampling procedure, which is novel for the dd-CRP. We evaluate on the TREC Twitter Timeline Generation (TTG), obtaining encouraging results: despite using a weak baseline retrieval model, the dd-CRP story clustering method is competitive with the best entries in the 2014 TTG task.

1 Introduction

A long-standing goal for information retrieval and extraction is to identify and group textual references to ongoing events in the world (Allan, 2002). Success on this task would have applications in personalized news portals (Gabrilovich et al., 2004), intelligence analysis, disaster relief (Vieweg et al., 2010), and in understanding the properties of the news cycle (Leskovec et al., 2009). This task attains a new importance in the era of social media, where citizen journalists can document events as they unfold (Lotan et al., 2011), but where repetition and untrustworthy information can make the reader’s task especially challenging (Becker et al., 2011; Marcus et al., 2011; Petrović et al., 2010).

A major technical challenge is in fusing information from two heterogeneous data sources: textual content and time. Two different documents about a single event might use very different vocabulary, particularly in sparse social media data such as microblogs; conversely, two different sporting events might be described in nearly identical language, with differences only in the numerical outcome. Temporal information is therefore critical: in the first case, to find the commonalities across disparate writing styles, and in the second case, to identify the differences. A further challenge is that unlike in standard document clustering tasks, the number of events in a data stream is typically unknown in advance. Finally, there is a high premium on scalability, since online text is produced at a high rate.

Due to these challenges, existing approaches for combining these modalities have been somewhat heuristic, relying on tunable parameters to control the tradeoff between textual and temporal similarity. In contrast, the Bayesian setting provides elegant formalisms for reasoning about latent structures (e.g., events) and their stochastically-generated realizations across text and time. In this paper, we describe one such model, based on the distance-dependent Chinese Restaurant Process (dd-CRP; Blei and Frazier, 2011). This model is distinguished by the neat separation that it draws between textual content, which is treated as a stochastic emission from an unknown Multinomial distribution, and time, which is modeled as a prior on graphs over documents, through an arbitrary *distance function*. However, straightforward implementations of the dd-CRP are insufficiently scalable, and so the model

has been relatively underutilized in the NLP literature (Titov and Klementiev, 2011; Kim and Oh, 2011; Sirts et al., 2014). We describe improvements to Bayesian inference that make the application of this model feasible, and present encouraging empirical results on the Tweet Timeline Generation task from TREC 2014 (Lin et al., 2014).

2 Model

The basic task that we address is to group short text documents into an unknown number of storylines, based on their textual content and their temporal signature. The textual content may be extremely sparse — the typical Tweet is on the order of ten words long — so leveraging temporal information is crucial. Moreover, the temporal signal is multi-scale: in the 24-hour news cycle, some storylines last for less than an hour, while others, like the disappearance of the Malaysian Airlines 370 plane in 2014, continue for weeks or months. In some cases, the temporal distribution of references to a storyline will be unimodal and well-described by a parametric model (Marcus et al., 2011); in other cases, it may be irregular, with bursts of activity followed by periods of silence (He et al., 2007). Finally, it will be crucial to produce an implementation that scales to large corpora.

The distance-dependent Chinese Restaurant Process (dd-CRP) meets many of these criteria (Blei and Frazier, 2011). In this model, the key idea is that each instance (document) i “follows” another instance c_i (where it is possible that $c_i = i$), inducing a graph. We can compute a partitioning over instances by considering the connected components in the undirected version of the follower graph; these partitions correspond to “tables” in the conventional “Chinese Restaurant” analogy (Aldous, 1985), or to clusters. The advantage of this approach is that it is fundamentally non-parametric, and it introduces a clean separation between the textual data and the covariates: the text is generated by a distribution associated with the partition, while the covariates are associated with the following links, which are conditioned on a distance function.

The distribution over follower links for document

i has the following form,

$$\Pr(c_i = j) \propto \begin{cases} f(d_{i,j}), & i \neq j \\ \alpha, & i = j, \end{cases} \quad (1)$$

where $d_{i,j}$ is the distance between units i and j , and $\alpha > 0$ is a parameter of the model. Large values of α induce more self-links and therefore more fine-grained partitionings. Since we are concerned with temporal covariates, we define the distance function as follows:

$$f(d_{i,j}) = e^{\frac{-|t_i - t_j|}{a}}. \quad (2)$$

Thus, the likelihood of document i following document j decreases exponentially as the time gap $|t_i - t_j|$ increases.

The text of each document i is represented by a vector of word counts \mathbf{w}_i . The likelihood distribution is multinomial, conditioned on a parameter θ associated with the partition to which document i belongs. By placing a Dirichlet prior on θ , we can analytically integrate it out. Writing $\mathbf{z}^{(c)}$ for the cluster membership induced by the follower graph \mathbf{c} , we have:

$$\begin{aligned} P(\mathbf{w} \mid \mathbf{c}; \eta) &= \prod_k P(\{\mathbf{w}_i : \mathbf{z}_i^{(c)} = k\}; \eta) \\ &= \prod_k \int_{\theta} P(\{\mathbf{w}_i : \mathbf{z}_i^{(c)} = k\} \mid \theta) P(\theta; \eta) d\theta \end{aligned} \quad (3)$$

$$(4)$$

Given a multinomial likelihood $P(\mathbf{w} \mid \theta)$ and a (symmetric) Dirichlet prior $P(\theta \mid \eta)$, this integral has a closed-form solution as the Dirichlet-Multinomial distribution (also known as the multivariate Polya distribution). The joint probability is therefore equal to the product of Equation 1 and Equation 4,

$$P(\mathbf{w}, \mathbf{c}) = \prod_i P(c_i; \alpha, a) \prod_k P(\{\mathbf{w}_i : \mathbf{z}_i^{(c)} = k\}; \eta). \quad (5)$$

The model has three hyperparameters: α , which controls the likelihood of self-linking, and therefore affects the number of clusters; a , which controls the time scale of the distance function, and therefore affects the importance of the temporal dimension to the resulting clusters; and η , which controls the precision of the Dirichlet prior, and therefore the importance of rare words in the textual likelihood function.

Estimation of these hyperparameters is described in § 3.2.

3 Inference

The key sampling equation for the dd-CRP is the posterior likelihood,

$$\Pr(c_i = j \mid \mathbf{c}_{-i}, \mathbf{w}) \propto \Pr(c_i = j)P(\mathbf{w} \mid \mathbf{c}).$$

The prior is defined in Equation 1. Let ℓ represent the likelihood under the partitioning induced when the link c_i is cut. Now, the likelihood term has two cases: in the first case, j is already in the same connected component as i (even after cutting the link c_i), so no components are merged by setting $c_i = j$. In this case, the likelihood $P(\mathbf{w} \mid \mathbf{c}_i = j)$ is exactly equal to ℓ . In the second case, setting $c_i = j$ causes two clusters to be merged. This gives the likelihood,

$$P(\mathbf{w} \mid c_i = j, \mathbf{c}_{-i}) \propto \frac{P(\{\mathbf{w}_k : z_k^{(c)} = z_j^{(c)} \vee z_k^{(c)} = z_i^{(c)}\})}{P(\{\mathbf{w}_k : z_k^{(c)} = z_i^{(c)}\})P(\{\mathbf{w}_k : z_k^{(c)} = z_j^{(c)}\})},$$

where the constant of proportionality is exactly equal to ℓ . Each of the terms in the likelihood ratio is a Dirichlet Compound Multinomial likelihood. This likelihood function is itself a ratio of gamma functions; by eliminating constant terms and exploiting the identity $\Gamma(x + 1) = x\Gamma(x)$, we can reduce the number of Gamma function evaluations required to compute this ratio to the number of words which appear in *both* clusters $z_i^{(c)}$ and $z_j^{(c)}$. Words that occur in neither cluster can safely be ignored, and the gamma functions for words which occur in exactly one of the two clusters cancel in the numerator and denominator of the ratio. Note also that we only need compute the likelihood for c_i with respect to each cluster, not for every possible follower link.

3.1 Online inference

While we make every effort to accelerate the computation of individual Gibbs samples, the complexity of the basic algorithm is superlinear in the number of instances. This is due to the fact that each sample requires computing the probability of instance i joining every possible cluster, while the number of clusters itself grows with the number of instances

(this growth is logarithmic in the Chinese Restaurant Process). Scalability to the streaming setting therefore requires more aggressive optimizations.

To get back to linear time complexity, we employ a fixed-lag sampling procedure (Doucet et al., 2000). After receiving instance i , we perform Gibbs sampling only within the fixed window $[t_i - \tau, t_i]$, leaving c_j fixed if $t_j < t_i - \tau$. This approximate sampling procedure implicitly changes the underlying model, because there is no possibility of linking i to a later message j if the time gap $t_j - t_i > \tau$.

Since we are only interested in obtaining a single storyline clustering — rather than a full Bayesian distribution over clusterings — we perform annealing for samples towards the end of the sampling window. Specifically, we set the temperature to $\gamma = 2.0$ and exponentiate the sampling likelihood by the inverse temperature (Geman and Geman, 1984). This has the effect of interpolating between probabilistically-correct Gibbs sampling and a hard coordinate-ascent procedure.

3.2 Hyperparameter estimation

The model has three parameters to estimate:

- α , the concentration parameter of the dd-CRP
- a , the offset of the distance function
- η , the scale of the symmetric Dirichlet prior.

We interleave maximization-based updates to these parameters with sampling, in a procedure inspired by Monte Carlo Expectation Maximization (Wei and Tanner, 1990). Specifically, we compute gradients on the likelihood $P(\mathbf{c})$ with respect to α and a , and take gradient steps after every fixed number of samples. For the symmetric Dirichlet parameter η , we employ the heuristic from Minka (2012) by setting the parameter to $\eta = \frac{(K-1)/2}{\sum_k \log p_k}$, where K is the number of words that appear exactly once, and p_k is the probability of choosing the k^{th} word from the vocabulary under the unigram distribution for the entire corpus.

4 TREC Evaluation

To test the efficacy of this approach, we evaluate on the Twitter Timeline Generation (TTG) task in the Microblog track of TREC 2014. It involves taking tweets based on a query Q at time T and returning

a summary that captures relevant information. We perform the task on 55 queries with different timestamps and compare our results with 13 groups that submitted 50 runs for this task in 2014.

We consider the following systems:

Baseline We replace the distance-dependent prior with a standard Dirichlet prior. The number of clusters is heuristically set to 20. Annealed Gibbs sampling is employed for inference.

Offline inference The dd-CRP model with offline inference procedure (described in § 3).

Online inference The dd-CRP model with online inference procedure (described in § 3.1).

For the online inference implementation, we set the size of window and number of iterations to five days and 500 respectively. For the baseline, the parameter of the Dirichlet prior was set to a vector of 0.5 for each cluster. These values were chosen through 10-fold cross validation.

To measure the quality of the clusterings obtained by these models, we compare the average weighted and unweighted F-measures for 55 TREC topics, using the evaluation scripts from the TREC TTG task. Overall results are shown in Table 1. The ONLINE MODEL has the best weighted F1 score, outperforming the offline version of the same model, even though its inference procedure is an approximation to the OFFLINE MODEL. It may be that its approximate inference procedure discourages long-range linkages, thus placing a greater emphasis on the temporal dimension. Both models were trained over 500 iterations, and the ONLINE MODEL was 30% faster to train than the offline model.

Compared to the other 2014 TREC TTG systems, our dd-CRP models are competitive. Both models outperform all but one of the fourteen submissions on the unweighted F_1 metric, and would have placed fourth on the weighted F_1^w metric. Note that the TREC evaluation scores both clustering quality and retrieval. We use only the baseline retrieval model, which achieved a mean average precision of 0.31. The competing systems shown in Table 1 all use retrieval models that are far superior: the retrieval model for top-ranked PKUICST team (line 4) achieved a mean average precision (MAP) of 0.59 (Lv et al., 2014), and the QCRI (Magdy et al.,

2014) and and hltcoe (Xu et al., 2014) teams (lines 5 and 6) used retrieval models with MAP scores of at least 0.5. Bayesian dd-CRP storyline clustering was competitive with these timeline generation systems despite employing a far worse retrieval model, so improving the retrieval model to achieve parity with these alternative systems seems the most straightforward path towards better overall performance.

5 Related work

Topic tracking and first-story detection are very well-studied tasks; space does not permit a complete analysis of the related work, but see (Allan, 2002) for a summary of “first generation” research. More recent non-Bayesian approaches have focused on string overlap (Suen et al., 2013), submodular optimization (Shahaf et al., 2012), and locality-sensitive hashing (Petrović et al., 2010). In Bayesian storyline analysis, the seminal models are Topics-Over-Time (Wang and McCallum, 2006), which associates a parametric distribution over time with each topic (Ihler et al., 2006), and the Dynamic Topic Model (Blei and Lafferty, 2006), which models topic evolution as a linear dynamical system (Nallapati et al., 2007). Later work by Diao et al. (2012) offers a model for identifying “bursty” topics, with inference requiring dynamic programming. All these approaches require the number of topics to be identified in advance. Kim and Oh (2011) apply a distance-dependent Chinese Restaurant *Franchise* for temporal topic modeling; they evaluate using predictive likelihood rather than comparing against ground truth, and do not consider online inference.

The Infinite Topic-Cluster model (Ahmed et al., 2011a) is non-parametric over the number of storylines, through the use of the recurrent Chinese Restaurant Process (rCRP). The model is substantially more complex than our approach. Unlike the dd-CRP, the rCRP is Markovian in nature, so that the topic distribution at each point in time is conditioned on the previous epoch (or, at best, the previous K epochs, with complexity of inference increasing with K). This Markovian assumption creates probabilistic dependencies between the topic assignment for a given document and the documents that follow in subsequent epochs, necessitating an inference procedure that combines sequential

| Model | Rec. | Rec. ^w | Prec. | F ₁ | F ₁ ^w |
|---------------------------------------|------|-------------------|-------|----------------|-----------------------------|
| <i>dd-CRP clustering models</i> | | | | | |
| 1. BASELINE | 0.14 | 0.27 | 0.33 | 0.20 | 0.30 |
| 2. OFFLINE | 0.32 | 0.47 | 0.27 | 0.29 | 0.34 |
| 3. ONLINE | 0.34 | 0.55 | 0.26 | 0.29 | 0.35 |
| <i>Top systems from Trec-2014 TTG</i> | | | | | |
| 4. TTGPKUICST2 (Lv et al., 2014) | 0.37 | 0.58 | 0.46 | 0.35 | 0.46 |
| 5. EM50 (Magdy et al., 2014) | 0.29 | 0.48 | 0.42 | 0.25 | 0.38 |
| 6. hltcoeTTG1 (Xu et al., 2014) | 0.40 | 0.59 | 0.34 | 0.28 | 0.37 |

Table 1: Performance of Models in the TREC 2014 TTG Task. Weighted recall and F_1 are indicated as Rec.^w and F_1^w .

Monte Carlo and Metropolis Hastings, and a custom data structure; this inference procedure was complex enough to warrant a companion paper (Ahmed et al., 2011b). The rCRP is also employed by Diao and Jiang (2013, 2014). In contrast, the dd-CRP makes no Markovian assumptions, and efficient inference is possible through relatively straightforward Gibbs sampling in a fixed window.

6 Conclusion

We present a simple non-parametric model for clustering short documents (such as tweets) into storylines, which are conceptually coherent and temporally focused. Future work may consider learning more flexible temporal distance functions, which could potentially represent temporal periodicity or parametric models of content popularity.

Acknowledgments We thank the reviewers for their helpful feedback. This research was supported by an award from the National Institutes for Health (R01GM112697-01), and by Google, through a Focused Research Award for Computational Journalism.

References

- Ahmed et al., 2011a Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J. Smola, and Choon H. Teo. 2011a. Unified analysis of streaming news. In *WWW*, pages 267–276, Hyderabad, India.
- Ahmed et al., 2011b Amr Ahmed, Qirong Ho, Choon H Teo, Jacob Eisenstein, Eric P Xing, and Alex J Smola. 2011b. Online inference for the infinite topic-cluster model: Storylines from streaming text. In *AISTATS*, pages 101–109, Fort Lauderdale, FL.
- Aldous, 1985 David J Aldous. 1985. *Exchangeability and related topics*. Springer.
- Allan, 2002 James Allan. 2002. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media.
- Becker et al., 2011 Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 438–441.
- Blei and Frazier, 2011 David M Blei and Peter I Frazier. 2011. Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488.
- Blei and Lafferty, 2006 David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.
- Diao and Jiang, 2013 Qiming Diao and Jing Jiang. 2013. A unified model for topics, events and users on twitter. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- Diao and Jiang, 2014 Qiming Diao and Jing Jiang. 2014. Recurrent chinese restaurant process with a duration-based discount for event identification from twitter. In *The 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (SDM’14)*.
- Diao et al., 2012 Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 536–544, Jeju, Korea.
- Doucet et al., 2000 Arnaud Doucet, Simon Godsill, and Christophe Andrieu. 2000. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208.
- Gabrilovich et al., 2004 Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. 2004. Newsjunkie: providing personalized newsfeeds via analysis of information

- novelty. In *Proceedings of the 13th international conference on World Wide Web*, pages 482–490. ACM.
- Geman and Geman, 1984 Stuart Geman and Donald Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741.
- He et al., 2007 Qi He, Kuiyu Chang, and Ee-Peng Lim. 2007. Analyzing feature trajectories for event detection. In *SIGIR*, pages 207–214. ACM.
- Ihler et al., 2006 Alexander Ihler, Jon Hutchinson, and Padhraic Smyth. 2006. Adaptive event detection with time-varying poisson processes. In *KDD*, pages 207–216. ACM.
- Kim and Oh, 2011 Dongwoo Kim and Alice Oh. 2011. Accounting for data dependencies within a hierarchical dirichlet process mixture model. In *cikm*, pages 873–878.
- Leskovec et al., 2009 Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pages 497–506.
- Lin et al., 2014 Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. 2014. Overview of the trec-2014 microblog track. In *Proceedings of the Twenty-Third Text REtrieval Conference*.
- Lotan et al., 2011 Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and danah boyd. 2011. The arab spring—the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, 5:31.
- Lv et al., 2014 Chao Lv, Feifan Fan, Runwei Qiang, Yue Fei, and Jianwu Yang. 2014. PKUICST at TREC 2014 Microblog Track: feature extraction for effective microblog search and adaptive clustering algorithms for TTG. Technical report, DTIC Document.
- Magdy et al., 2014 Walid Magdy, Wei Gao, Tarek Elganainy, and Zhongyu Wei. 2014. Qcri at trec 2014: applying the kiss principle for the ttg task in the microblog track. Technical report, DTIC Document.
- Marcus et al., 2011 Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *chi*, pages 227–236. ACM.
- Minka, 2012 Thomas Minka. 2012. Estimating a dirichlet distribution. <http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf>.
- Nallapati et al., 2007 Ramesh M Nallapati, Susan Dittmore, John D Lafferty, and Kin Ung. 2007. Multiscale topic tomography. In *KDD*, pages 520–529. ACM.
- Petrović et al., 2010 Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 181–189, Los Angeles, CA.
- Shahaf et al., 2012 Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. 2012. Trains of thought: Generating information maps. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 899–908, Lyon, France. ACM.
- Sirts et al., 2014 Kairit Sirts, Jacob Eisenstein, Micha Elsner, and Sharon Goldwater. 2014. Pos induction with distributional and morphological information using a distance-dependent chinese restaurant process. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, MD.
- Suen et al., 2013 Caroline Suen, Sandy Huang, Chantat Eksombatchai, Rok Sasic, and Jure Leskovec. 2013. Nifty: a system for large scale information flow tracking and clustering. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 1237–1248.
- Titov and Klementiev, 2011 Ivan Titov and Alexandre Klementiev. 2011. A bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1445–1455. Association for Computational Linguistics.
- Vieweg et al., 2010 Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10*, pages 1079–1088, New York, NY, USA. ACM.
- Wang and McCallum, 2006 Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM.
- Wei and Tanner, 1990 Greg CG Wei and Martin A Tanner. 1990. A monte carlo implementation of the em algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.
- Xu et al., 2014 Tan Xu, Paul McNamee, and Douglas W Oard. 2014. Hltcoe at trec 2014: Microblog and clinical decision support.